DOCUMENT RESUME

ED 305 400 TM 012 968

AUTHOR Lunz, Mary E.; And Others
TITLE Equating Practical Examinations.
PUB DATE Mar 89
NOTE 30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Difficulty Level; *Equated Scores; Latent Trait Theory; *Licensing Examinations (Professions); Medicine
IDENTIFIERS Facet Analysis; Histology; Log Linear Models; Multifactorial Models; *Practical Tests; Rasch Model; *Test Equivalence

ABSTRACT

This paper describes and illustrates a method for equating examinations with multiple facets (i.e., items, examinees, judges, tasks, and rating scales). The data are from the practical section of two histotechnology certification examinations. The first practical examination involved 210 examinees, 14 judges, 15 slides, 3 tasks, and 2 rating scales. The second examination involved 301 examinees, 18 judges, 15 slides, 3 tasks, and 2 rating scales. The Rasch model is extended in the facets model to accommodate all facets of the practical examination. The elements of each facet are calibrated to a common log-linear scale. The variables defined for each facet are found to be definable and invariant across examinee groups. The principles of common element equating are then applied in the multiple facet situation. Two examinations are calibrated and the necessary common elements included. The items and the judges are found to be slightly easier or more lenient on the second examination. To equate examinations and secure equivalent standards, the scale of the second examination is adjusted to match the scale of the first examination. Four data tables, one figure, and five graphs are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED305400

Equating Practical Examinations

Mary E. Lunz
American Society of Clinical Pathologists

Benjamin D. Wright
University of Chicago

John Stahl
American Society of Clinical Pathologists

John M. Linacre
University of Chicago

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARY E. LUNZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM012968

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA., March 1989

Equating Practical Examinations

This paper describes and illustrates a method for equating examinations with multiple facets (e.g. items, examinees, judges, tasks and rating scales). The data are from the practical section of a certification examination. The Rasch model is extended in the facets model to accommodate all facets of the practical examination. The elements of each facet are calibrated to a common log-linear scale. The variables defined for each facet are found to be definable and invariant across examinee groups. The principles of common element equating are then applied in the multiple facet situation. Two examinations are calibrated and the necessary common elements included. The items and the judges are found to be slightly easier or more lenient on the second examination. To equate the examinations and secure equivalent standards, the scale of the second examination is adjusted to match the scale of the first examination.

Key words: Rasch model, Judge severity, item difficulty, IRT, Latent Trait, Test equating, Facets Model

Equating Practical Examinations

The purpose of equating is to place all examinations on the same scale so that equivalent performance standards can be implemented. The type of equating with which this study is concerned is horizontal equating. Horizontal equating assumes that two examinee groups are similar and that alternate test forms are intended to measure the same level of ability. Equating removes sampling differences and puts the items and examinees on the same scale (Baker, 1984). In principle, the measures derived after equating will have become equivalent so that it is possible to compare the performance of examinees who have taken different tests (Angoff, 1971).

Equating techniques for dichotomous multiple choice examinations have been studied in some detail (Skaggs and Lissitz, 1981; Lord and Wingersky, 1984; Wright, 1977, Engelhart and Asberg, 1983). Rasch model equating techniques have been found to be practical (Rentz and Bashaw, 1977) and robust even when all assumptions are not met (Forsyth, 1981).

The purpose of this paper is to explore a method for equating examinations which involve multiple facets such as judges, tasks and rating scales in addition to items and examinees, using an expansion of the Rasch model. To equate examinations with multiple facets, a measurement model which models the probability of a response as a monotonically increasing function of examinee ability, given the characteristics of the intermediating facets, is required. The most able examinee should have the highest ability measure regardless of the characteristics of the intermediating facets.

The Rasch model (1960/1980) estimates item difficulty and examinee ability independently of one another on a common linear scale. The probability (P_{ni}) of person n with ability B achieving a correct response on

item i with difficulty D is modeled as $\log((P_{ni})/(1-P_{ni})) = (B_n - D_i)$. This probability is governed by the difference between the ability of the person and the difficulty of the item (Wright and Stone, 1979).

The rating scale model, an expansion of the Rasch model, includes an additional function for the effect of a rating scale. The probability of person n with ability B achieving rating step x with difficulty F on item i with difficulty D is modeled as $\log((P_{nix})/(P_{nix-1})) = (B_n - D_i - F_x)$ (Wright and Masters, 1982).

The facets model, a further expansion of the Rasch model, adds a function for judge severity and/or any other facet that may be part of the examination. The probability of person n with ability B achieving rating step x on item i with difficulty D from judge J with severity C is modeled as $\log((P_{nijx})/(P_{nijx-1})) = (B_n - D_i - C_j - F_x)$ (Linacre, 1987).

With the facets model it is possible to construct a variable composed of the elements included in each facet of an examination so that a continuum from most to least within each facet is observable. The ordering of the elements within each facet provides the frame of reference for verifying the intended examination definition. Facets analysis enables the calibration of items essays, protocols, cases or tasks, etc. from most to least difficult and judges from most to least severe. Variation in judge severity has been documented in several studies (Cason and Cason, 1984; de Gruijter, 1984). Variation in the elements of the other facets has been generally disregarded.

OBJECTIVES

The objectives of this paper are twofold.

1) To demonstrate that useful variables for each facet of of a judge intermediated examination can be constructed; and that the calibrations of the elements within each facet are invariant.

2) To demonstrate that examinations with multiple facets can be equated using Rasch common element equating techniques.

Previous research has confirmed that variables for each facet of an examination can be constructed (Lunz, Linacre and Wright, 1988; Lunz and Stahl, 1989). The next requirement for equating is to demonstrate the invariance of item and judge calibrations across diverse examinee groups and the invariance of the examinee ability measures and judge severity calibrations across different examination forms. If the calibrations are invariant, then Rasch model equating techniques are feasible for examinations with multiple facets.

For examinations with two facets (item difficulty and person ability) item difficulty is anchored by including a set of common items on alternate examination forms. The mean difference in the difficulty of the common items measures the difference in difficulty between the two tests. This difference can be used to transform the ability scale of the second exam to that of the first exam.

When more than two facets are involved in the examination, it becomes necessary to anchor more than one facet in order to identify differences in difficulty between exams. The difference in the difficulty of the common elements in each facet measures the difference in the difficulty between the

two tests (see Figure 1). The facets model specifies that differences in difficulty are additive across all facets. Therefore, the sum of the differences in difficulty across the pertinent facets transforms the ability scale of the second test to that of the first anchor test. This transformation involves determining a common origin for the two scales since the two scales are modeled to have the same unit of measurement.

INSERT FIGURE 1 ABOUT HERE

METHODS

Data

Data are from two histotechnology practical examinations. The first practical examination involves 210 examinees, 14 judges, 15 slides, 3 tasks and 2 rating scales. The second examination involves 301 examinees, 18 judges, (8 of whom were common to both exams), 15 slides, (8 of which were common to both examinations), the same 3 tasks and the same 2 rating scales. The directions require the examinees to prepare the slides to specifications for each type of tissue and tissue stain. Examinees are provided with the criteria for grading.

Five facets are defined in these data. The facets for each examination are the same; however, the elements of three of the facets (items, judges, examinees) vary across examinations.

FACET One: (examinee ability, Bn) the ability measures for 210 examinees on the anchor exam and 301 different examinees on the second examination.

FACET Two: (task difficulty, Am) the three tasks on which all microscopic slides are judged on both examinations. The tasks are (1) processing: the preparation of tissue, (2) microtomy: the cutting of tissue and (3) staining: the coloring of tissue.

FACET Three: (difficulty of the item, Di) the fifteen microscopic slides prepared by the examinee to detailed specifications for tissue type, size, stain, and thickness. The examinations contain common and noncommon slides.

FACET Four: (severity of the judge, Cj) the judges who grade an examination. The number of judges varies with the number of examinations to be graded. Some judges grade for many years and are common elements to both examinations, while others are less experienced or new. All judges receive 2-3 hours of training prior to grading an examination.

FACET Five: (step difficulty of the rating scale, Fmk) the rating alternatives for processing: 1=acceptable and 0=unacceptable; microtomy: 1=acceptable and 0=unacceptable; staining: 3=above average, 2=average, 1=below average, 0=poor/unacceptable. The rating scales remain the same for all examinations. It has been verified that all judges use this scale similarly from examination to examination.

Grading Network

Judges grade examinations during two day grading sessions. Two to three hours of training occur prior to each grading session. During training, the expectations for each slide are reviewed, and the grading scales are discussed. The judges agree on grading criteria for each of the three tasks on each of the 15 slides. The tasks and grading scale are the same for all practical examinations, although the slides and judges may differ.

The construction of measurement objectivity for examinations with multiple facets requires that judges grade overlapping (but not necessarily complete) samples of examinees and slides. When this is laid out properly, it becomes possible to calculate relative item (D_i) and task (A_m) difficulties across all judges and relative judge severities (C_j) across all items and tasks.

The grading design used to obtain the necessary overlap specified that: (1) judges be randomly assigned to tables, (2) subsets of slides be rotated among tables and (3) no judge grade more than one subset of slides from an examinee. Each judge was seated at a table with a microscope and space to work independently. The 15 slides were divided into 3 subsets of 5 slides. Each table graded a particular subset of five slides during a specified period. In the next period each judge graded a different subset of slides so that each judge graded each of the 15 slides sometime during the two day session. A judge graded all three tasks for each slide and approximately ten examinees during each grading period. This produced about 150 observations (3 tasks per slide X 5 slides X 10 examinees) per judge, per period. This grading network produces complete overlap in the final data.

Analysis

FACETS (Linacre, 1988) a computer program for the Rasch analysis of examinations with multiple facets is used to analyze data from these two histotechnology practical examinations. FACETS places each element within each facet of the examination on a common scale according to the following probability equation.

$$\log\left[\frac{Pnmijk}{Pnmijk-1}\right] = Bn - Am - Di - Cj - Fmk$$

Where Pnmijk = probability of person n being given grade k by judge j on item i of task m
Pnmijk-1 = probability of person n being given grade k-1 by judge j on item i of task m
Bn = ABILITY of person n
Am = DIFFICULTY of task m
Di = DIFFICULTY of item i
Cj = SEVERITY of judge j
Fmk = DIFFICULTY of grading step k on task m relative to step k-1

The FACETS program produces three statistics for each element of each facet: a logit calibration, its standard error of estimate, and a model fit assessment. In addition, a mean, standard deviation and separation reliability (comparable to the KR-20) are calculated for the elements within each facet.

Invariance Study

Rasch measurement requires that person measures be statistically equivalent regardless of the items appearing on a test, and that item and judge calibrations be statistically equivalent regardless of the particular examinee population. These requirements must be met before equating examinations with multiple facets is possible.

To study the invariance of the facets calibrations, the second (301 examinees) histotechnology practical is analyzed under two conditions. To assess the invariance of the calibrations for items and judges, the examinee population is divided into thirds for the first condition. The facets for items and judges are calibrated twice. The lower third of the examinee population is used for the first calibration and the upper third of the examinee population is used for the second calibration. For invariance, the item and judge calibrations should be statistically equivalent across these diverse examinee groups (Wright, 1967).

In the second condition, the item free characteristic of the examinee ability measures and the judge severity calibrations are assessed. The slides from the second histotechnology practical are divided into two test forms (1 and 2) with common slides between forms. This resembles the method used to construct examinations and is therefore a good test of the process. Due to

the grading network all judges grade all slides on both forms. Examinee ability measures are calculated separately from each form. Judge severity calibrations are also calculated from each form. For invariance, examinee ability measures and judge severity calibrations should be statistically equivalent across forms (Wright, 1967).

Equating Study

Both practical examinations are analyzed using FACETS. This statistically defines the variables for the facets in each examination. Common elements are used to equate the exams. The facets with elements which vary between examinations are slides (items) and judges (see Figure 1). Therefore, differences in slide difficulty and judge severity must be accounted for in the equating process. Since common elements exist on the variables for slides and judges, it is possible to measure the between-test difference in difficulty of each facet. The additive structure of the facets model specifies that the cumulative difference of the two facets is the amount by which the second exam must be adjusted to place it on the same scale as the first anchor exam.

RESULTS

Invariance Study Results

In the first condition of the invariance study the calibrations for slides and judges are expected to be statistically equivalent even though the examinees have diverse characteristics. Tables 1 and 2 and Graphs 1 and 2 report the results when slides and judges are calibrated against the most able

and least able examinee groups. On Table 1 the standardized difference shows that slide 8 is significantly harder for the lower third of the population and slide 14 is significantly harder for the upper third of the population. The other 13 slides maintain their position on the scale and could be used for test equating.

On Table 2 the standardized difference shows judge 3 is less severe when grading the upper third and judges 12, 15 and 16 are less severe when grading the lower third. The effect of these judge inconsistencies on examinee ability measures is minimized by the grading network. The other 14 judges maintain their relative severity across the diverse examinee populations, indicating that most judges maintain their level of severity regardless of the examinee group they are grading. These 14 judges could be used for equating.

INSERT TABLES 1 & 2 AND GRAPHS 1 & 2 ABOUT HERE

In the second condition of the invariance study, judge severity calibrations and examinee ability measures are expected to be statistically equivalent across test forms. Form 1 was found to be .02 logits more difficult than form 2. Graph 3 shows that judge 10 graded significantly more leniently on form 2, but that the other judges maintain a comparable level of severity across test forms.

Graph 4 shows the comparison of examinee ability measures from form 1 and form 2. The correlation is .84 and when corrected for attenuation becomes .99. A few examinees performed differently on forms 1 and 2. To provide a frame of reference, each pair of examinee measures is compared by forming

their standardized difference. The examinee measures on the forms are expected to be statistically equivalent. On graph 5, the standardized differences are plotted against examinee ability measures. Only 3, out of 301 examinees show significantly different measures on form 1 and form 2. Further study of these examinees, 165, 31, and 14, revealed the following. Examinee 14 was significantly less able on form 2 and examinee 31 was significantly less able on form 1 because each had submitted one wrong tissue for a slide. This resulted in a score of zero (0) on that slide which lowered their ability measures significantly on that form. Examinee 165 earned a near perfect score on form 2 but had one entirely unacceptable slide on form 1. This aberration significantly affected the ability measures between forms. The measures of examinee 160 were not quite significantly different, but did show the same pattern as examinee 165. The other examinee ability measures are not significantly different given the error of the measure. The mean examinee ability measures on forms 1 and 2 are 1.06 logits (SD .77, SE .38) and .94 logits (SD .78, SE .40) respectively. These results show that common element equating is feasible for examinations with multiple facets.

INSERT GRAPHS 3, 4, 5 ABOUT HERE

Equating Study Results

All facets of Exam 1 are calibrated and Exam 1 serves as the anchor scale. The difficulties of the 15 slides range from is 1.00 to -.31 logits with a mean of .00 and SD of .34 logits (see Table 3). The separation reliability (comparable to the KR-20) is .94. These data show that a clear

continuum from easy to difficult is defined by these slides. The severities of the 14 judges range from .60 to -.43 logits with a mean of .00 and SD of .28 logits (see Table 4). This separation reliability is .81. These data show a significant variation in judge severity that must be accounted for to achieve objectivity and fairness.

All facets of Exam 2 are also calibrated (see Tables 3 & 4). The 15 slide difficulties on Exam 2 range from 1.16 to -.57 logits with a mean of .00 and SD of.39 logits. This separation reliability is .97. A clear continuum of easy to difficult is also defined by these slides. The severities of the 18 judges range from .46 to -.61 logits with a mean of .00 and SD of .28 logits. This separation reliability is .92. These data also show significant variation in judge severity. Since the tasks and grading scales are known to be the same for both examinations; the slides and judges are the source of difference between the two exams.

Common slides and common judges appear on both examinations. This facilitates the implementation of common element equating. Tables 3 and 4 report the calibrations and standard errors for the slides and judges common to both exams as well as the non-common elements. Table 3 shows that the mean difference in difficulty on the common slides is .00 logits; however, the standardized difference for slide 6 shows that it is significantly easier on the Exam 2 (the amount is .26 logits with a standard error of .12 logits). Table 4 shows that the mean difference in the severity of the common judges is .10 logits. The standardized difference shows that judges 2 and 6 grade significantly more severely on Exam 2 while judges 4 and 8 grade significantly more leniently on Exam 2. The effect of judge grading bias on examinee

measures is minimized within each exam. However, judge inconsistency between exams is sufficient reason to eliminate a judge as a common element in the equating.

INSERT TABLES 3 & 4 ABOUT HERE

The next step is to control the quality of the links between the two examinations. This is done by dropping slide 6 and judges 2, 4, 6, and 8 from the equating because the standardized difference indicates significant variance in their performance across examinations. The mean difference in the difficulty of the remaining slides becomes -.03 logits (SD .10) and these slides are consistent across examinations. The mean difference in the severity of the remaining judges becomes -.07 logits (SD .10) and these judges are consistent across examinations.

The effect of the facets is additive, so the difference between Exam 1 (anchor) and Exam 2 is -.10 logits, (-.03) + (-.07) = (-.10). The second examination is slightly easier than the first because the slides are slightly easier and the judges are slightly more lenient. To become comparable to Exam 1, the examinee measures for Exam 2 must be adjusted by -.10 logits.

Exam 2 is anchored to Exam 1 using the 8 slides and 5 judges found to be consistent across examinations. Before anchoring the mean examinee ability measure on Exam 2 is 1.10 logits (SD .71). After the anchoring, the mean examinee ability measure on Exam 2 is adjusted down to 0.99 logits (SD .71) because of the slightly easier slides and the slightly less severe judges on

Exam 2. The scale of the second exam is adjusted so that the measure scales of Exams 1 and 2 become comparable and the same pass/fail standard can be implemented consistently across both examinations.

DISCUSSION

In the past, parallel examinations with multiple facets were assumed equivalent. The validity of this assumption was doubtful but there was little choice but to "believe" in it. The facets model provides a psychometric method for assessing each element within each facet, regardless of the number of facets required for an examination. It is no longer necessary to accept equivalence on faith, the tools are available to demonstrate the possibility of equivalence and to construct and maintain it.

This study demonstrates that the elements within each facet vary significantly, and that common element equating can be applied to identify and correct for differences within facets between examinations. While the techniques will certainly benefit from more research and refinement, these results are encouraging and provide more information than has been available. In particular judges vary substantially in their level of severity, but tend to be consistent across examinee groups. Practical examination items also vary in difficulty, so the implementation of standards based only on local analysis are unlikely to be consistent from exam to exam. Since training seems to improve inter or intrajudge consistency minimally, accounting psychometrically for the differences between exams is the only way to control examination differences. Further research regarding the performance and selection of facets for equating is underway.

The fit of the data to the model was not discussed, although these data did generally fit. The interpretation of fit in the context of an examination with multiple facets must be explored. The effect of measurement error must also be explored in the context of analyzing and equating examinations with multiple facets. The calibrations were found to be stable across examinee groups in this study; however, more research into the robustness and reproductability of the calibrations is in order.

Achieving psychometric control over examinations with multiple facets opens many possibilities for achieving measurement objectivity for all judge intermediated evaluations. It should be possible to use the full potential of a practical, oral or essay examination without the distraction of the subjective noise which can invalidate the process. Equating examinations with multiple facets initiates the opportunity to control for differences in the difficulty of the parallel examinations and to insure that standards are applied consistently across examinations.

References

Angoff, W.H. (1971). Scales, norms and equivalent scores in R.L. Thorndike (ed.) Educational Measurement, Washington, D.C.: American Council on Education, 508-600.

Baker, F.B. (1984). Ability metric transformations involved in vertical equating under item response theory. Applied Psychological Measurement, 8, 261-271.

Cason, G.J. and Cason, C.L. (1984). A deterministic theory of clinical performance rating. Evaluation and the Health Professions, 7, 221-247.

Engelhart, G. and Asberg, D.W. (1983). Constructing a test network with a Rasch Measurement model. Applied Psychological Measurement, 7, 3, 283-294.

Forsyth, R., Sarsangjan, V., Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model, Applied Psychological Measurement 5, 175-186.

de Gruijter, D.N.M. (1984). Two simple models for rater effects. Applied Psychological Measurement, 8, 2, 213-218.

Linacre, J.M. (1987). An Extension of the Rasch Model to Multi-faceted Situations. Chicago; University of Chicago.

Linacre, J.M. (1988). FACETS, Chicago, IL: MESA Press.

Lord, F.M. & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". Applied Psychological Measurement. 8, 453-461.

Lunz, M.E., Linacre, J.M. & Wright, B.D. (1988). The impact of judge severity on examination scores. Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA.

Lunz, M.E., Stahl, J.A., (1989). Variations among examinees and practicals on oral examinations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Rasch, G. (1960/1980). Probablistic models from some intelligence and attainment tests. Chicago: University of Chicago Press.

Rentz, L.R. & Bashaw, W.L. (1977). The national reference scale for reading: an application of the Rasch model. Journal of Educational Measurement, 14, 161-178.

Skaggs, G. & Lissitz, R. (1982). Test equating: relevant issues and a review of recent research. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Wright, B.D. & Masters, G.N. (1982). Rating Scale Analysis. Chicago: MESA Press.

Wright, B.D. & Stone, M.H. (1979). Best Test Design. Chicago: MESA Press.

Wright, B.D. (1967). Sample-free test calibration and person measurement. Paper presented at the Invitational Conference on Testing Problems, Educational Testing Service, Princeton, N.J.

Wright, B.D. (1977). Solving measurement problems with the Rasch Model. Journal of Educational Measurement, 14, 97-116.

Table 1
Invariance Study: Condition 1
Slide Difficulty Calibrations Invariance
Upper 1/3 Compared to Lower 1/3 Examinees

SLIDE NUMBER	UPPER THIRD DIF	UPPER THIRD SE	LOWER THIRD DIF	LOWER THIRD SE	DIFFER*	SE DIFF**	STANDARDIZED DIFF***
1	0.33	0.13	0.33	0.09	0.00	0.16	0.00
2	-0.20	0.15	-0.08	0.10	-0.12	0.18	-0.67
3	-0.41	0.17	-0.13	0.10	-0.28	0.20	-1.42
4	0.10	0.14	0.15	0.09	-0.05	0.17	-0.30
5	-0.02	0.15	0.06	0.10	-0.08	0.18	-0.44
6	-0.10	0.15	-0.39	0.11	0.29	0.19	1.56
7	-0.19	0.15	-0.10	0.10	-0.09	0.18	-0.50
8	0.01	0.14	0.43	0.09	-0.42	0.17	**-2.52**
9	-0.08	0.15	0.17	0.10	-0.25	0.18	-1.39
10	1.07	0.10	1.11	0.08	-0.04	0.13	-0.31
11	-0.46	0.18	-0.57	0.11	0.11	0.21	0.52
12	-0.13	0.16	-0.12	0.10	-0.01	0.19	-0.05
13	-0.06	0.15	-0.09	0.10	0.03	0.18	0.17
14	0.27	0.14	-0.39	0.11	0.66	0.18	**3.71**
15	-0.13	0.16	-0.38	0.11	0.25	0.19	1.29
MEAN	0.00	0.15	0.00	0.10	0.00	0.18	-0.02
S.D.	0.35	0.02	0.40	0.01	0.25	0.02	1.40****

DIF= Difficulty calibration in logits
* Upper 1/3 DIF - Lower 1/3 DIF
** $((\text{Upper Third SE})^2 + (\text{Lower Third SE})^{1/2}$
*** DIFFER/SEDIFF (criteria = ± 2.00)
**** Expected to be 1.00
13 of the 15 slides suitable for equating (slides 8 and 14 not suitable)
After removing slides 8 and 14, the mean standardized difference is -.01 and the SD is .84.

Table 2
Invariance Study: Condition 1
Judge Severity Calibrations Invariance
Upper 1/3 Compared to Lower 1/3 Examinees

JUDGE NUMBER	UPPER THIRD SEVERITY	UPPER THIRD SE	LOWER THIRD SEVERITY	LOWER THIRD SE	DIFFER*	SE DIFF**	STANDARDIZED DIFFER***
1	0.09	0.15	0.13	0.12	-0.04	0.19	-0.21
2	-0.05	0.14	-0.09	0.11	0.04	0.18	0.22
3	-0.82	0.16	-0.41	0.10	-0.41	0.19	**-2.17**
4	0.18	0.12	-0.02	0.11	0.20	0.16	1.23
5	-0.22	0.16	0.14	0.10	-0.36	0.19	-1.91
6	-0.02	0.20	0.03	0.13	-0.05	0.24	-0.21
7	0.38	0.19	0.50	0.11	-0.12	0.22	-0.55
8	0.22	0.16	0.27	0.10	-0.05	0.19	-0.26
9	-0.42	0.26	-0.14	0.11	-0.28	0.28	-0.99
10	0.11	0.16	0.25	0.10	-0.14	0.19	-0.74
11	0.20	0.17	0.14	0.15	0.06	0.23	0.26
12	0.02	0.12	-0.47	0.12	0.49	0.17	**2.89**
13	-0.13	0.21	-0.25	0.12	0.12	0.24	0.50
14	-0.01	0.15	0.20	0.11	-0.21	0.19	-1.13
15	0.79	0.16	0.39	0.11	0.40	0.19	**2.06**
16	0.52	0.19	0.08	0.11	0.44	0.22	**2.00**
17	-0.84	0.19	-0.65	0.14	-0.19	0.24	-0.81
18	0.01	0.14	-0.08	0.12	0.09	0.18	0.49
MEAN	0.00	0.17	0.00	0.12	-0.00	0.20	0.04
S.D.	0.39	0.03	0.29	0.01	0.25	0.03	1.31****

Severity = Severity calibrations in logits
* Upper 1/3 Severity - Lower 1/3 Severity
** $((\text{Upper Third SE})^2 + (\text{Lower Third SE})^{1/2}$
*** DIFFER/SEDIFF (criteria = ± 2.00)
**** Expected to be 1.0
14 of 18 judges suitable for equating (judges 3, 12, 15, 16 not suitable)
After removing judges 3, 12, 15, 16, the mean standardized difference is .05 and the SD .79.

Table 3
Equating Exam 2 to Exam 1
Slide Difficulties including Common Slides

Exam 1			Exam 2					
Slide Item #	Difficulty Calib	SE	Slide Item #	Difficulty Calib	SE	DIFFER*	SE DIFF**	Standardized DIFFER***
COMMON SLIDES								
1	.24	.08	1	.38	.06	-.14	.10	-1.40
2	-.01	.08	2	-.07	.07	.06	.11	.56
3	-.22	.09	3	-.19	.03	-.03	.11	.26
4	.27	.08	4	.15	.06	.12	.10	1.20
5	1.00	.07	5	1.16	.05	-.16	.09	-1.86
6	-.31	.09	6	-.57	.08	.26	.12	**2.16 @**
7	-.26	.09	7	-.31	.07	.05	.11	.44
8	-.23	.09	8	-.10	.07	-.13	.11	-1.14
Mean	.60	.08		.60	.07	.00		-.04
SD	.41	.01		.49	.01	.14		1.29****
NON-COMMON SLIDES								
9	.21	.08						
10	-.16	.09						
11	-.18	.09						
12	-.30	.09						
13	-.15	.09						
14	.26	.08						
15	-.16	.09						
			9	.01	.07			
			10	-.37	.07			
			11	.30	.06			
			12	.06	.07			
			13	-.12	.07			
			14	-.14	.07			
			15	-.31	.07			

* Difficulty Exam 1 - Difficulty Exam 2
** $((SE_1)^2 + (SE_2)^2)^{1/2}$
*** DIFFER/SEDIFF
**** Expected = 1.00
@ Slide six was not included as common for equating

Table 4
Equating Exam 2 to Exam 1
Judge Severity Including Common Judges

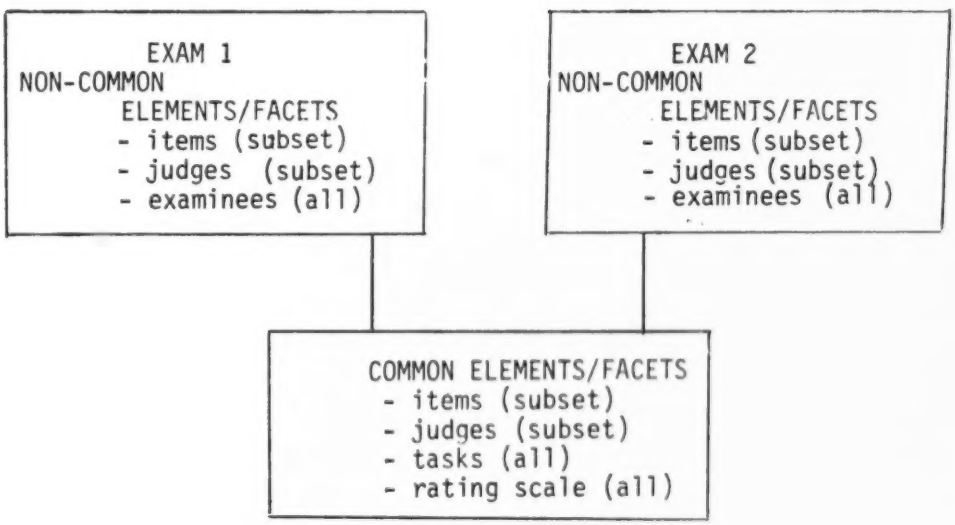
Exam 1			Exam 2						
Judge #	Severity Calib	SE	Judge #	Severity Calib	SE	DIFFER*	Scale Center	SE DIFF**	Standard DIFFER***
COMMON JUDGES									
1	.31	.07	1	.10	.07	.21	.11	.10	1.11
2	-.28	.08	2	.02	.06	-.30	-.40	.10	**-4.00** @
3	-.01	.09	3	-.04	.07	.03	-.07	.11	- .61
4	.60	.08	4	-.04	.08	.64	.54	.11	**4.77** @
5	-.19	.10	5	-.14	.09	-.05	-.15	.13	-1.11
6	.11	.08	6	.38	.07	-.27	-.37	.11	**-3.48** @
7	.22	.09	7	.23	.08	-.01	-.11	.12	- .91
8	.09	.07	8	-.61	.09	.70	.60	.11	**5.26** @
9	-.13	.09	9	-.07	.08	-.06	-.16	.12	-1.33
Mean	.08	.08		-.02	.08	.10	.00		-.03
SD	.26	.01		.26	.01	.34	.34		3.06****
NON-COMMON JUDGES									
10	.10	.08							
11	-.37	.06							
12	.18	.34							
13	-.21	.09							
14	-.43	.09							
			10	-.51	.07				
			11	.33	.08				
			12	-.24	.09				
			13	.15	.07				
			14	.18	.09				
			15	-.13	.06				
			16	-.29	.09				
			17	.12	.07				
			18	.46	.07				

* Severity Exam 1 - Severity Exam 2
** $((SE_1)^2 + (SE_2)^2)^{1/2}$
*** DIFFER/SEDIFF
**** Expected = 1.00
@ These judges were not included as common for equating

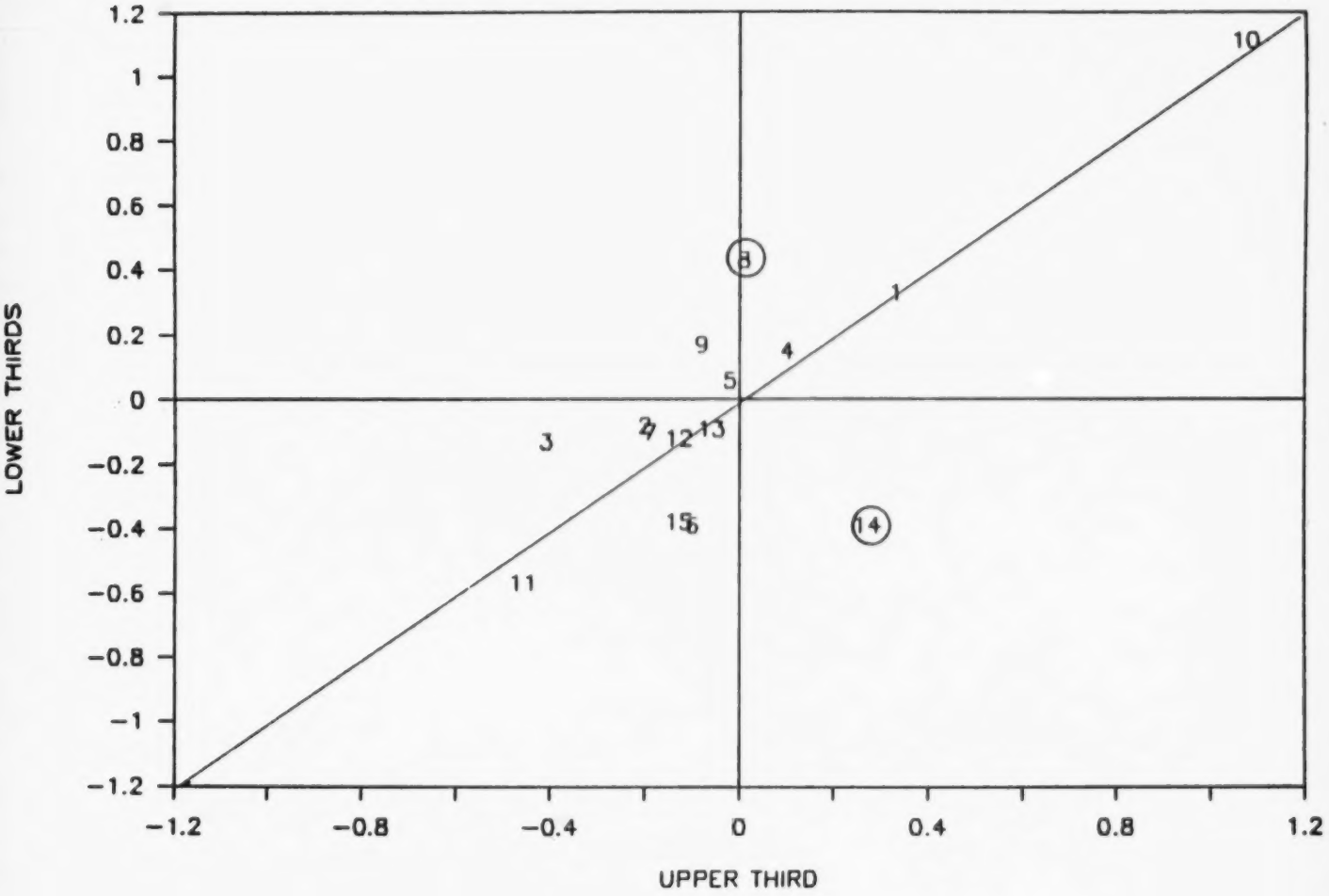
FIGURE 1

COMMON LINKS FOR EQUATING EXAMINATIONS

WITH MULTIPLE FACETS

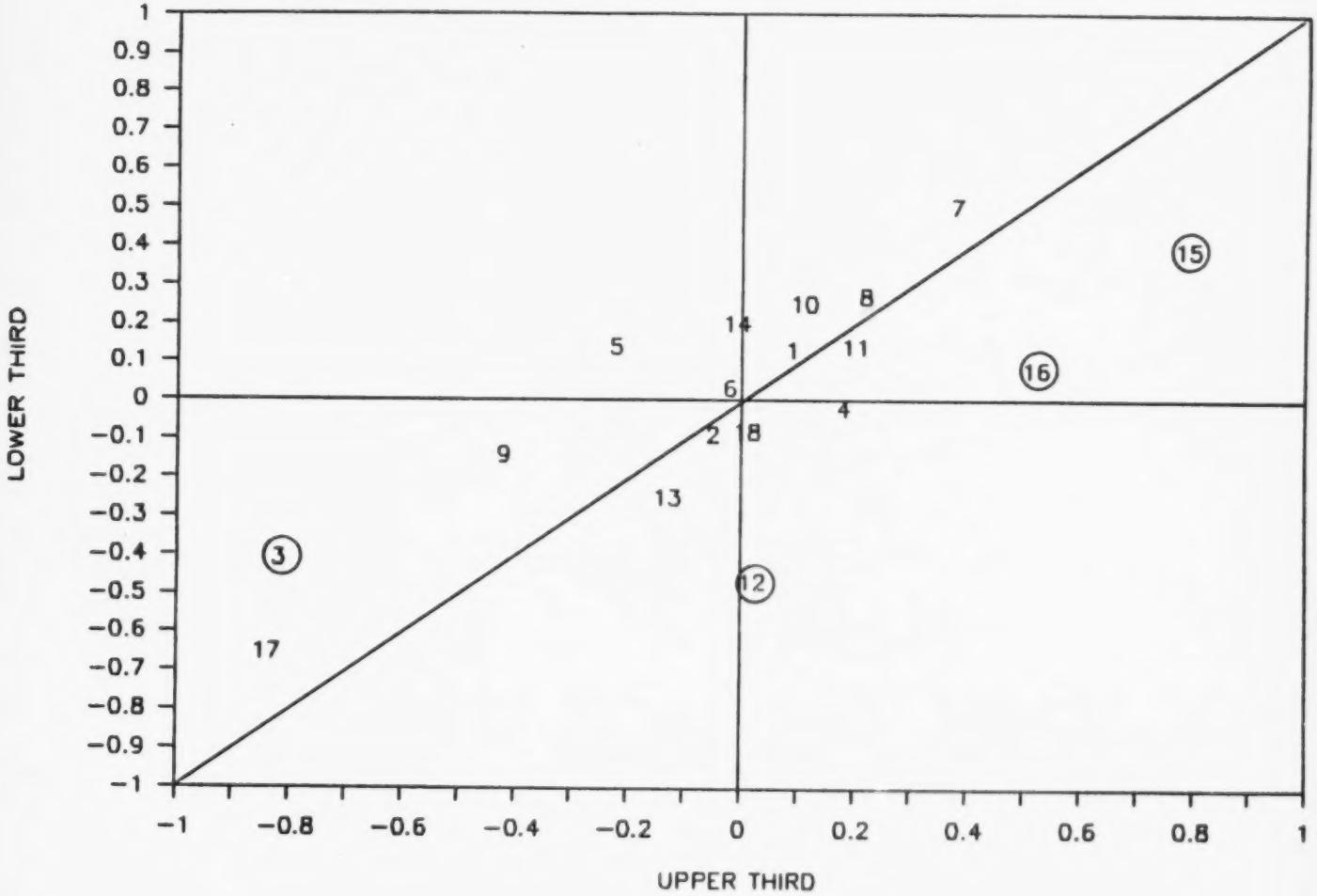


GRAPH 1
INVARIANCE STUDY: CONDITION 1
SLIDE DIFFICULTY CALIBRATIONS FOR UPPER AND LOWER THIRDS



GRAPH 2
INVARIANCE STUDY: CONDITION 1

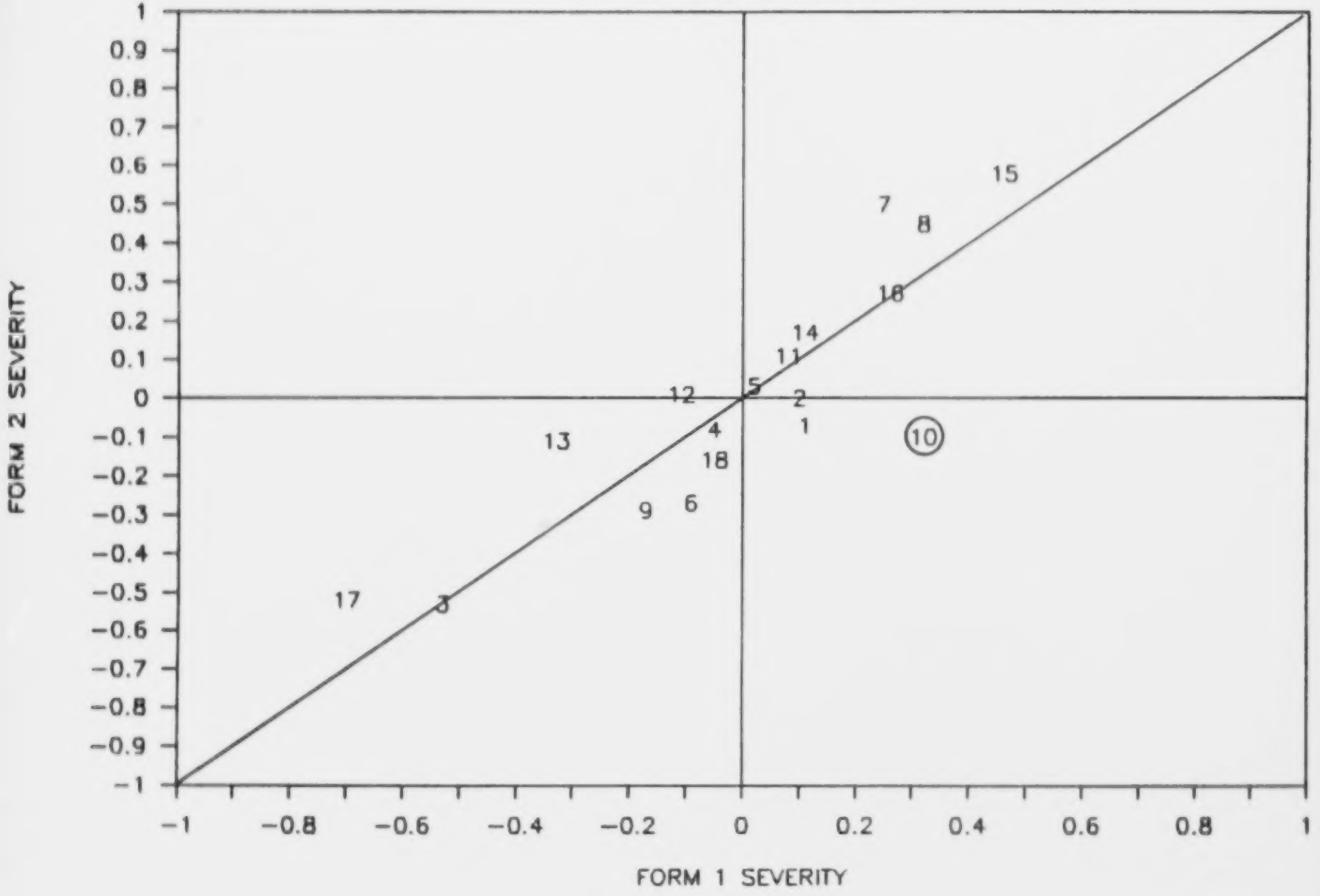
JUDGE SEVERITY CALIBRATIONS FOR UPPER AND LOWER THIRDS



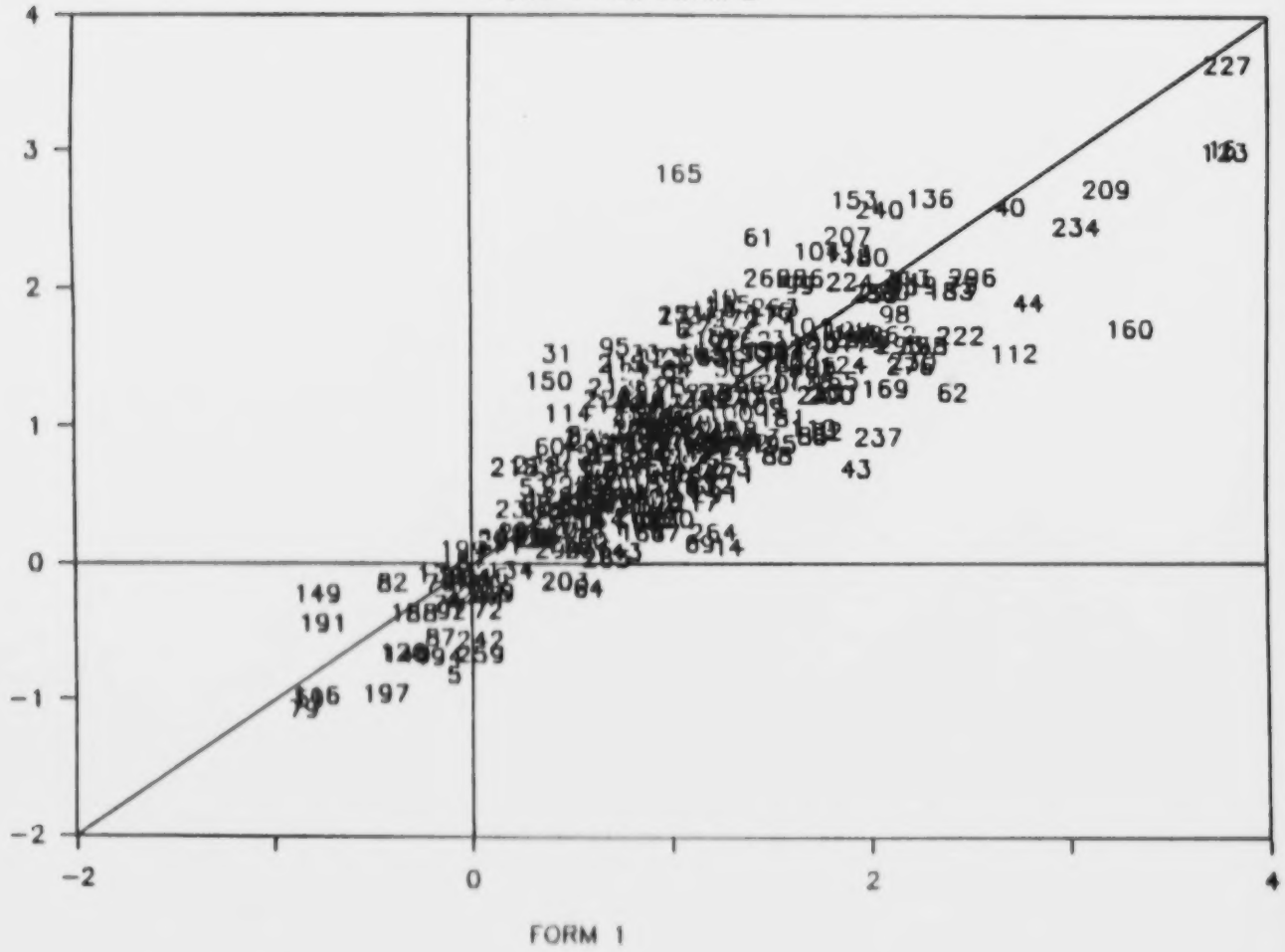
GRAPH 3

INVARIANCE STUDY: CONDITION 2
JUDGE SEVERITY CALIBRATIONS ACROSS

FORM 1 AND FORM 2



GRAPH 4
INVARIANCE STUDY: CONDITION 2
EXAMINEE ABILITY MEASURES ACROSS
FORM 1 AND FORM 2

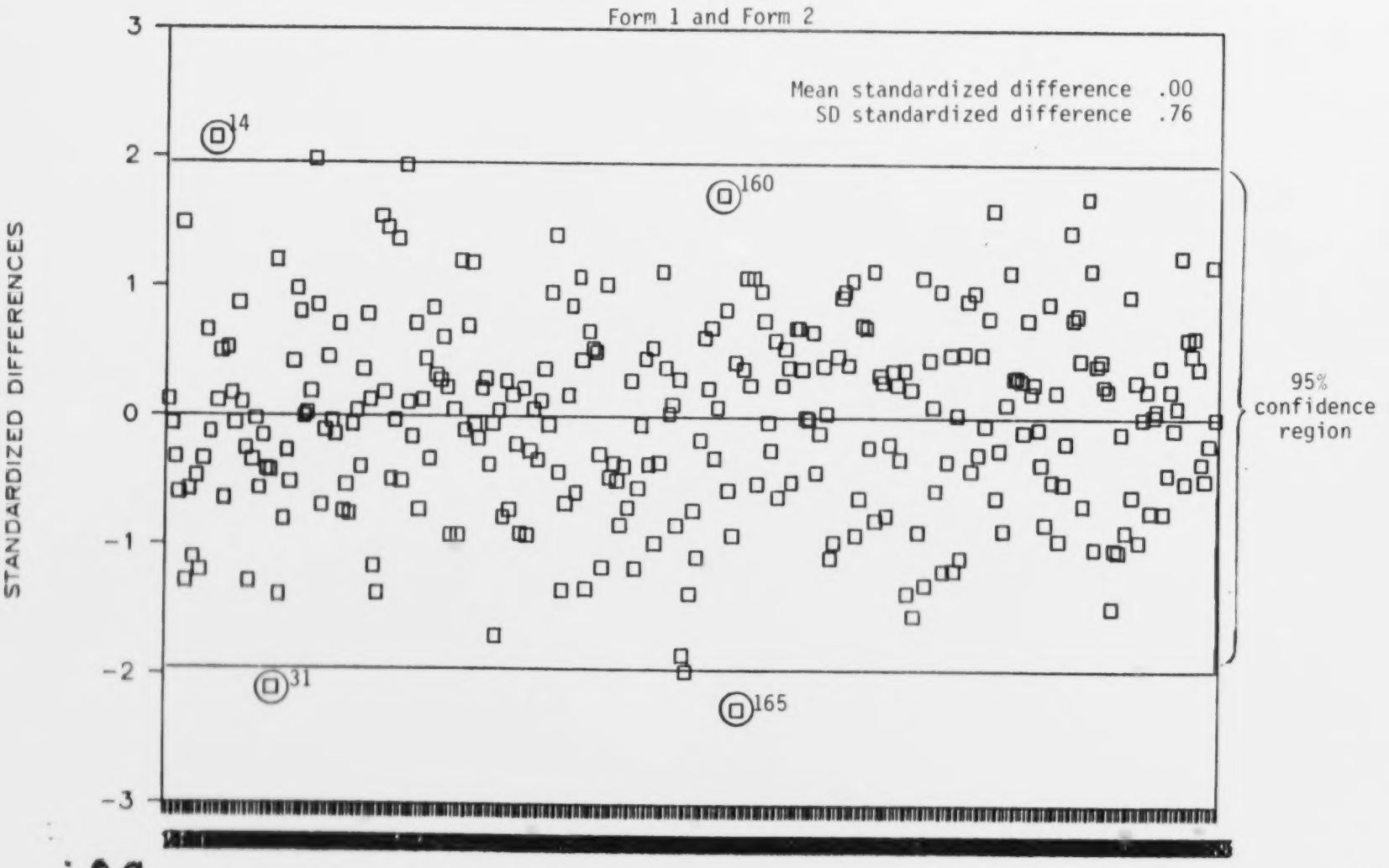


N = 301 examinees
Form 1: mean = 1.06 logits (SD .77)
Form 2: mean = .94 logits (SD .78)

GRAPH 5

INVARIANCE STUDY: CONDITION 2

COMPARISON OF EXAMINEES ABILITY MEASURES ACROSS
STANDARDIZED DIFFERENCES IN ABILITY



N = 310 examinees

95% confidence region indicates scores did not change given the error of measurement